



Taken from *Health Affairs*, July/August 1998, pp. 7-25.

## Performance Measurement: Problems And Solutions

Measuring performance of health plans remains elusive, but not from lack of effort or brains. One expert outlines a strategy to improve our chance of success.

By David M. Eddy

**PROLOGUE:** The issues around the quality of medical care have become a subject of increasing complexity confusion, and public interest. As managed care has rapidly emerged as the preferred delivery and financing mode of purchasers, if not all providers, it has thrust upon physicians new demands for accountability. However, the available tools to accurately measure performance of providers are quite limited. In this lead paper David Eddy discusses the challenges facing individuals and organizations that are striving to create better ways to measure performance and--importantly--offers some constructive ways to address the shortcomings.

As a physician (with a degree from the University of Virginia) who operates at the intersection of medicine and applied mathematics (doctorate, Stanford University), Eddy is certainly well equipped for the task. Beyond his academic credits, Eddy also has devoted considerable energy to helping the National Committee for Quality Assurance (NCQA) pursue the complex tasks around improved measurement, including service on the Committee on Performance Measurement, which produces the Health Plan Employer Data and Information Set (HEDIS) measures. Eddy also sees performance measurement from the viewpoint of providers and plans. He has consulted widely with many medical organizations and currently serves as senior adviser for health policy and management at Kaiser Permanente Southern California. *Health Affairs* invited several responses to Eddy's paper, which was based on the Richard and Hinda Rosenthal Lecture that he presented recently at the Institute of Medicine, National Academy of Sciences.

**ABSTRACT:** Recent efforts to measure performance have established its feasibility and value. However, its full potential is currently limited by several problems. They include the probabilistic nature, rarity, and confounding of many health outcomes; the inadequacies of information systems; the multiplicity of measurers and measures; the complexity of health plans; and the availability or funding. Solutions are to rely more on process measures; to justify every measure with a formal, evidence-based rationale; to improve information systems; to supplement population-based measures with case-based measures; to develop a single, nationally standardized set of measures; and to provide

**nonpolitical, public funding for the design and administration of measures.**

Given the importance of health care, it seems inconceivable that we do not have excellent ways of evaluating how well we are doing. Yet the fact is, we do not. Our attempts to systematically measure the quality of care are less than a decade old and still very much in their methodological adolescence. The delay in getting started can be explained by a variety of factors: a general assumption that quality was high, the implied insult to the medical profession and discomfort to the public that comes with measuring performance, and the fact that substandard performance is largely invisible except through a statistical lens. It took exposés of poor quality and questions from purchasers about what they were getting for their money to push performance measurement ahead.

However, once it is launched, its importance can be profound. Performance measurement informs people of the outcomes they can expect from certain treatments; it is the basis of plans' efforts to improve care; and it is the basis for people's choices of providers and plans. Beyond that, performance measures drive health care resources in very specific directions. When a national organization such as the National Committee for Quality Assurance (NCQA) publishes a Health Plan Employer Data and Information Set (HEDIS) measure, the effect is as if every health plan in the country went on a retreat to set their clinical goals for the coming year and all came back with the same answer. I cannot think of a more powerful single instrument for shifting health care resources than a national set of performance measures would be.

Now that we appreciate the need for and Importance of performance measurement, it is reasonable to ask why, after a decade of work, we still have not been able to build the measurement tools we need. The reason is not lack of effort or brains, but the fact that performance measurement is just plain difficult. This paper describes some essential concepts, the main problems, and some solutions. Because my experience in performance measurement has been with an organization that develops measures (the NCQA) and a health plan that reports measures (Kaiser Permanente Southern California), my observations reflect those perspectives.

**Concepts**

The design of a performance measure, and therefore how good it is, depends on several factors: the purpose of the measure, the entity whose quality is being measured, the dimension of quality being measured, the type of measure, and who will use the measure. It is important to identify these, because a measure that is good for one purpose, entity, dimension, or audience might be bad for another.

**Purpose.** Performance measurement has three main purposes. The simplest purpose is to describe the effect of some intervention on a specified group of patients--as in a typical "outcomes study." This can be achieved with a single measurement taken after an appropriate follow-up time. It is the easiest purpose to design measures for, not only because it involves a single measurement but also because there are no comparisons being made across interventions, time, or surgeons, factors that might confuse such comparisons.

The next step up in difficulty is to measure an improvement in outcomes caused by some modification of a treatment or care process--as in a quality improvement program. This is more complex because it involves taking measurements at two times and requires that all other factors that might affect the outcomes remain unchanged during the interval between measurements. Despite this, the comparison is still relatively simple because it is being made within a single health plan. Although there might be a few changes in the plan or its population that could confound the comparison, their effects are usually small.

The third purpose is to compare the quality of care being delivered by different entities such as different health plans, medical groups, hospitals, or physicians (for simplicity, I call all of these "plans"). Making comparisons across plans is much more difficult, first, because the outcomes plans want to achieve, and (or which they are being held responsible, are only partially under their control. Second, there are many differences between plans and the populations they serve that can affect outcomes independently of the quality of care the plan is delivering.

A fourth purpose that needs to be considered is that someone may want to promote a measure simply to stimulate plans to give priority to some intervention, disease, or population. This type of pressure can come from an advocacy group, a disease-oriented charity, a specialty society, or a pharmaceutical company. Although such a measure is often cast as a desire to compare plans, that may in fact be its secondary purpose.

How good a measure is depends heavily on its purpose. One common error is to take a measure that was designed to track outcomes (purpose 1) and try to use it to compare plans (purpose 3). Confounding factors that are moot for the first purpose may destroy the measure for the second. Conversely, measures designed to compare plans may be far too blunt to track outcomes or measure improvement. Just because a measure has been widely used, "validated," successful, or intuitively right for one purpose does not necessarily mean that it will be right for another.

**Entity being measured.** The design of a measure also can be affected by the entity being measured. In addition to health plans, it is possible to evaluate such entities as hospitals, nursing homes, medical groups, specialty departments, and physicians. The important issues include the extent to which there is a defined population, the degree of control the entity has over the delivery of care, the logistics of tracking patients, the quality of data systems, and the number of patients available for observation (sample size). Again, a measure that is proven for one entity may not work for another.

**Dimension of quality.** A third important factor is the dimension of quality that is being assessed. The main dimensions are coverage, access, choice of provider, service (for example, hotel-style amenities, courtesy), information about health plans, and clinical care. Each of these requires different methods.

**Type of measure.** Most of the measures used to compare the quality of care delivered by various health plans are based on populations. "Population-based" measures begin with a group of people who are candidates for some intervention and calculate the proportion who have a particular outcome. In this context, the "outcome" could be the performance of the intervention, a biological outcome, or a health outcome (an outcome people can experience and care about). The first case is called a "process measure," because it measures some aspect of the process of care that was performed. The latter two are often called "outcome measures," although strictly speaking that term should be reserved for health outcomes. Examples of population-based measures are the proportion of patients with hypertension who are being treated (a process measure), the proportion of treated hypertensives whose pressures are controlled to below 140/90 mm Hg (a biological measure), and the proportion of hypertensives who have heart attacks (an outcome measure).

We prefer to measure health outcomes because they are what people really care about, they are comprehensible, they aggregate the effects of all of the things plans do for a condition, and they leave plans free to determine for themselves the best things to do. For example, while the proportion of women who receive mammograms measures the quality of a plan's outreach program, breast cancer deaths would measure not only the plan's outreach program but every other aspect of care right up to death: the calibration of the equipment, the accuracy of the radiologists' interpretation, the follow-up of positive results, and the quality of treatment.

**Intended audience.** This factor affects the appropriate level of detail and the clinical sophistication required to understand what a change or difference in a measure means. For example, a measure intended to help a physician evaluate the effects of an asthma treatment can be much more detailed and clinically oriented than a measure intended to help asthmatic patients choose an asthma clinic.

The remainder of this paper addresses the most difficult measurement problem: the use of measures to compare the clinical care of health plans, to help purchasers and consumers choose plans. The purpose (comparison), target (health plans), dimension (clinical outcomes), and audience (lay people) force us to think about all of the issues that can make performance measurement so difficult.

## Why Is It So Difficult?

The idea behind the population-based measure is both simple and powerful: There is a group of people who either have or are at high risk of getting some condition. Plans can either prevent the condition from occurring or improve its outcome. Therefore, to determine a plan's quality, we can identify the persons who are candidates for the desired interventions and measure the rates at which either the interventions are used or the desired outcomes occur. As straightforward as this might appear, many problems can make the use of population-based measures extremely difficult, especially for those based on health outcomes. These problems can be grouped into two main categories: "natural" and "man-made." The latter can possibly be changed, whereas the former can only be worked around.

**Natural problems.** There are six main natural problems.

*Probability factor.* The first problem is central: Almost all health outcomes are highly probabilistic. They do not always occur when a plan does the right thing, and they can occur even when a plan does the wrong thing. This makes the measurement of quality in health care fundamentally different from quality measurement in most other industries. Specifically, although we can see individual outcomes, such as "Mr. Smith" having a heart attack, we cannot draw the same types of conclusions that we would if we saw a toaster burn a slice of bread. Conclusions about the quality of health care require a large number of observations and statistical analysis.

*Low frequency.* The second natural problem, low frequency, compounds the first. The unfortunate fact is that many of the health outcomes that are most interesting do not occur very frequently. This is true not only for rare diseases but also for many of the most common and socially important diseases. For example, any measure of breast cancer care has to deal with the fact that its most important outcome—death—occurs at a rate of about one per 1,000 per year in women over age fifty. That by itself is a small effect to find. But the problem is made worse by the fact that performance measurement is looking not at that rate (death versus no death) but at the change in that rate caused by how various plans apply various interventions. For example, screening (versus no screen) might change a plan's breast cancer death rates by about three per 10,000 per year, and a difference in screening rates that might be seen in different plans (say, ten percentage points) might cause breast cancer death rates to differ by three per 100,000 women per year. The low frequency of outcomes has big implications for the sample size needed to measure a meaningful difference in outcomes across plans. If breast cancer mortality were to be used as a measure of breast cancer screening, a population of about two million women would be needed to find that size difference in mortality. The median-size health maintenance organization (HMO) has fewer than 10,000 women over age fifty, which makes this measure impossible to use for comparing the quality of breast cancer care.

*Long delays.* It takes five to ten years to get five-to-ten-year survival rates. This not only affects the logistics of measurement (for example, finding patients after five to ten years) but makes that measure of quality five to fifteen years out of date the day it is calculated.

*Control over outcomes.* The next natural problem concerns the extent to which a plan can control outcomes, versus the extent to which the outcomes are determined by other factors that are beyond a plan's control. A good example is low birth weight. For this to be a good measure for comparing the quality of prenatal care, ideally, all of the differences (variance) in the weights of babies should be explained by where the mother got her care. Unfortunately, an analysis of data from the HEDIS pilot project found that the plan the mother went to explained only about 5 percent of the variance in birth weight. Not only is this a very weak effect, it also means that there must be many other factors over which a plan has little or no control that have much (about twentyfold) stronger effects on birth weight. Some of these other factors are known and could possibly be affected by a plan (tobacco and alcohol use explain about 9

percent of the variance), others are known and could possibly be adjusted for (race explains about 5 percent more of the variance) but the majority (controlling about 80 percent of the variance) are not even known, much less under a plan's control or adjustable. Low birth weight may be an excellent measure for advocating attention to a vulnerable population, but it is a very poor measure for determining the actual quality of care plans deliver to that population.

*Level of clinical detail.* Because performance occurs through clinical decisions, measurement of it ideally should be done at that level of clinical detail. Just how detailed that is can be appreciated by studying the designs of clinical trials, because these provide the evidence for clinical decisions and have had to operationalize clinical terms the way performance measures must. Ideally, it should be possible to convert any clinical trial into a performance measure, and the level of detail required for a performance measure should be consonant with that of its corresponding clinical trial.

Consider, for example a recent drug trial that showed the value of adding a long-acting beta agonist for patients who have moderate asthma and persistent symptoms despite medium doses of inhaled glucocorticoids. Nationally prominent asthma guidelines recommended treatment with beta agonists, and it is certainly a reasonable candidate for a performance measure. To create such a measure, we first need an operational definition of "moderate asthma." The definition in the trial was patients "who have had asthma for at least six months and had been treated with an inhaled glucocorticoid for at least three months. The forced expiratory volume in 1 second (FEV) at baseline has to be at least 50% of the predicted value with an increase of at least 15% in FEV from the baseline after inhalation of 1 mg of terbutaline." Other terms that have similarly detailed operational definitions are "medium dose glucocorticoids," "persistent," symptoms, and "severe" attacks. Even if the definitions in the trials are softened, very detailed clinical data spanning patient characteristics, diagnoses, symptoms, test results, treatments, and outcomes clearly are needed.

*Comprehensibility.* To appreciate the last natural problem comprehensibility, recall that the primary objective of measurement is to help purchasers and consumers choose among plans. Population-based measures do this by reporting something like "the rate of (some outcome) was 70 percent in Plan A versus 76 percent in Plan B." The problem is that although many outcomes such as heart attacks or emergency room visits are understandable, many others are not. This is especially true of biological outcomes (for example, how important is a five-percentage-point difference in the proportion of asthmatics with FEV below 50 percent?) and process measures (for example, what difference does it make if 5 percent more moderate asthmatics are treated with beta agonists?) Indeed, few if any physicians could tell you what either of those means in terms of outcomes that patients really care about.

*Man-made problems.* If the natural problems are not bad enough, they are made worse by some man-made ones. To call them "man-made" is not to say that anyone in particular is to blame, but only to indicate that they follow from the way our health care system has evolved and, in theory at least, could be changed.

*Inadequate information systems.* The first such problem follows from the fact that performance measures require a high level of clinical detail. The unfortunate fact is that few health plans have sufficiently powerful information systems to obtain the required information. Most still rely on paper medical records. The computerized systems they have were implemented piecemeal to serve other, administrative, purposes. For performance measurement they are blunt and cumbersome. The only way to get the level of detail needed to convert a typical clinical trial into a performance measure is by a manual review of charts, which is extremely expensive. Today's information systems place an inherent limit on the quality of today's performance measures.

*Too many measurers and measures.* A second problem is the multiplicity of measurers and measures. Health plans today are bombarded by scores of different measurers who are demanding hundreds of different measures. Some, such as the NCQA's HEDIS, were designed by employers and plans together to standardize and simplify the burden on plans, but many others are put out by individual companies and consulting firms whose incentives are to splinter rather than to consolidate. The requested measures are rarely coordinated and are sometimes in conflict. Even when several measures appear to measure the same thing, seemingly small differences in definitions, time periods, sampling methods, or adjustment factors can create computational nightmares and destroy accurate comparisons. For example, if one measure of hypertension control asks for the proportion of known hypertensives whose blood

pressures are under 140/90 mm Hg, while another asks for the proportion whose systolic pressures have been reduced 20 percent below pretreatment levels, the result is two measures that look alike and invite comparison but in fact produce very different answers, even when applied to the same group of hypertensives. Thus, this problem hits two ways; It not only increases the burden on plans, but it also threatens the accuracy of comparisons.

*Health plan complexity.* A third man-made problem is the complexity of health plans. When a plan contracts with a wide network of physicians, each of whom is contracting with a wide range of plans, the result is measurement chaos. Setting aside the issues of how a plan might control the quality of care delivered by each provider and how providers might respond to various guidelines and quality-management programs from various plans, multiple, decentralized charts and small sample sizes cause problems. A related issue is that different purchasers require different benefit packages. Thus, a measure might ask about the performance of an intervention that is covered by some benefit packages but not by others.

*Funding.* A final man-made problem is funding. The economics of performance measurement are driven by the fact that many (but by no means all) large purchasers are requiring plans to provide measures of performance. Plans bear the cost of collecting the data and reporting the measures. In addition to their need to respond to purchasers' requests, plans may have a positive financial incentive if they do well on a set of measures and if advertising that fact motivates purchasers and consumers to choose them. Unfortunately, not all plans do well, and those that do poorly have a financial incentive to drop out. As more drop out, the performance bar moves up, and even fewer will do well on the next round of measures. Furthermore, the support from purchasers is thin. Many do not understand performance measurement, or cannot tell a good measure from a bad one, or are really more interested in cost than in quality. The financial incentives of those who design measures are either perverse or flat. Consulting firms that market their particular measures do get paid for their work, but their incentives are to create as many different measures as possible, which both burdens plans and confuses consumers. Nonprofit, quality-oriented organizations such as the NCQA have a strong philosophical motivation to do the work, but there is no market that enables them to capture the value or recover their costs. They rely on financial grants and gifts, but even those have to be screened lest they come from parties that hope a measure will enhance their own interests. The fact is that a publicly available, standardized set of performance measures is a public good, but there is no public money to support it. The funding of performance measurement is quite unstable, and unless there is a change, it is unlikely that this work can be sustained.

## **How Bad Is It?**

The net effect of all of these problems is that today's measures tend to be blunt, expensive, incomplete, and distorting. And, unless great care is taken, they can easily be inaccurate and misleading. The bluntness is a consequence of the poor information systems: "sharp" measures require a high level of clinical detail, which today's information systems cannot provide without massive spending. This greatly limits the number of good measures that are possible.

The expense of measurement is due to the multiplicity of measures and, when it is required, the cost of reviewing charts. Even when the data can be drawn from administrative systems, the cost of linking different data sets, cleaning the data, doing the calculations, and writing the reports is still quite high. An analysis of measures considered for HEDIS 3.0 for example, found that the cost to a plan ranged from \$20,000 to \$700,000 per measure.

The incompleteness is due to both the high costs and the natural factors that limit feasibility (such as the rarity or delay of outcomes). Out of the thousands of things plans do that affect quality, even a "comprehensive" (and expensive) measurement set can look at only a few dozen. Even when we think we have a measure for a condition, it usually measures only a small portion of the quality of care related to it. For example, the proportion of treated hypertensives whose blood pressures are controlled to 140/90 may be labeled a "hypertension measure," but it totally misses activities to prevent hypertension, screen people to find hypertensives, or ensure that known hypertensives are treated. It even misses the benefits of bringing someone's pressure down from 200/140 to 145/95.

As for inaccuracy, it can creep into performance measurement through every pore. Some of the most obvious sources are insufficient sample sizes, inaccuracies in data sets, and the

presence confounding factors that are either understood but not adjusted for or not understood at all. Seemingly small details in a measure's specifications can create biases that favor one plan over another or distort incentives. For example, defining the quality of hypertension treatment in terms of the proportion of pressures reduced to below 140/90 creates a very different therapeutic goal than asking about the proportion of pressures lowered to 20 percent of the pretreatment level. At the extreme, with the former definition, a plan could ignore patients with extremely high pressures that could not be brought below 140/90 without hurting their performance on the measure. Of all of these sources of inaccuracy, the only one now' included in reported measures is the effect of sample size.

That measures can be misleading follows from the facts that they are incomplete (We are trying to determine how well a plan does hundreds of things by looking at about twenty of them) and can be inaccurate. But before we leave this topic, we need to appreciate that this is an especially pernicious and dangerous problem, because it can be virtually impossible for anyone who looks at a measure to determine how accurate it is. An absolutely terrible measure will still produce a result, which for all intents and purposes will look just as authentic as the result produced by an accurate measure. The only visible feature of a measure is the confidence interval calculated from the sample size. Errors stemming from long delays, data systems, poor specification, or confounding factors will be invisible.

Measures can be distorting because health plans quite naturally want to do as well as they can with interventions and outcomes that are being measured and thus will put more resources into them. Although this is a desirable and often intentional consequence of measurement, it can yield misleading results: The quality of things that are measured may be higher than the quality of things that are not. It also can have the undesirable consequence of causing plans to spend fewer resources on other activities that are just as important but not being measured. Even worse, if a measure happens to address a relatively less important, less effective, or inefficient activity, by siphoning off resources the measure can actually have a detrimental effect on the quality of care.

Obviously, these problems are harmful to purchasers and consumers. But they are particularly frustrating to plans that are being required to spend precious measurement resources measuring things they know to be poorly understood, inaccurate, and misleading.

## Solutions

Fortunately, several things can be done to help solve these problems.

**Process measures.** The problems of probabilism, rarity, delay, weak control, and confounding place a fundamental limit on the extent to which measures can be based on health outcomes. As much homage as we pay to health outcomes, we simply cannot force them to do things they cannot do. Specifically, when the main health outcomes for an important condition are infrequent, delayed, weakly controllable, or heavily confounded, blind adherence to outcomes will produce inaccurate results. Inaccurate results are not just a statistical problem. They cause patients, physicians, and plans to make bad decisions. A poorly designed outcome measure can easily do more harm than good.

The solution is to use more process measures. Unlike many of their companion health outcomes, processes tend to be frequent, immediate, controllable, and rarely confounded by other factors. If properly designed process measures also can also steer plans toward particular activities that are known to be effective. Some process measures, especially those involving "elective" processes such as cancer screening, do depend on patients' compliance. But for most processes, compliance either is not an issue (for example, thrombolytics for heart attacks) or can be assumed to be similar across plans.

Process measures do have some drawbacks. First, they require an assumption that a difference in the process represents an important difference in health outcomes. Second, a change in a process measure is less meaningful and more easily misunderstood than is a health outcome. Most people understand what it means to prevent a breast cancer death, but few understand

what it means if one plan's mammography screening rate is 10 percent higher than another's. The third is that a process measure, by its very nature, micromanages. Instead of leaving plans free to set their own priorities for improving health outcomes, a process measure tells plans precisely what their priorities should be.

Biological outcomes also can play a role as proxies for health outcomes. One strength is that biological outcomes, like health outcomes, leave plans free to choose their own interventions. But biological outcomes vary widely in their vulnerability to the problems that can harm health outcomes. For example, counting the proportion of early-stage breast cancers as a measure of screening effectiveness corrects for the long delay of breast cancer deaths but is still far too infrequent to be feasible (requiring about 1.5 million women to find a difference between plans) and is heavily confounded by the possibility of overdiagnosis and lead-time bias. Blood pressure is both immediate and common, making its measurement preferable to counting heart attacks, but it introduces other problems. For example, a person's blood pressure fluctuates widely, its measurement depends heavily on the circumstances and technique, and it can be confounded by factors beyond a plan's control. Furthermore, like most biological measures, it is a continuous measure that must be manipulated in some way to make it useful (for example, by defining a threshold for "control" such as 140/90 mm Hg). Such manipulations are inevitably artificial and oversimplified and can create the types of distortions already described. In short, when there is a good correlation between a biological outcome and an important health outcome, the biological outcome should be considered, but a great deal of thought must be given to its properties.

**Formal workup.** Before any measure is promoted, there should be a formal analysis of its clinical significance, statistical characteristics, relevance, feasibility, and cost-effectiveness. Particularly important themes are that measures should be evidence based, calibrated, and cost-effective. There are two main reasons to require this. The first is to ensure that any measure is valid--that it measures what it purports to measure, accurately. The second is to ensure that plans are not burdened with unimportant inaccurate, misleading, or superfluous measures.

*Clinical significance.* Understanding a measure's clinical significance begins with a requirement that there be solid evidence, preferably from randomized trials, documenting that there is a causal link between a plan's actions and the desired outcomes. Every health outcome measure should have evidence that plans can affect it, and every biological measure should be supported by evidence both that represents or predicts an improvement in health outcomes and that there are things plans can do to affect it. When such evidence exists, we say that a measure is evidence based.

The requirement for supporting trials not only solves the first drawback of process and biological measures (connections to health outcomes) but also helps to solve the second drawback (comprehensibility) by enabling a quantitative estimate of how a change in a process or biological outcome changes a health outcome. For example, from trials of mammography screening it is possible to say that in an HMO that has 10,000 women between the ages of fifty and seventy-five, a 10 percent difference in mammography rates represents about one fewer breast cancer death every three years. A five-percentage-point difference in average FEV for moderate asthmatics represents one fewer severe attack per patient every four years. This idea can be summarized in the concept of calibration. Before any measure is promoted to compare plans, there should be a calculation of what a standardized change in the measure represents in terms of changes in the health outcomes of real interest.

A final thing that clinical analysis should do is ensure that the measure's specifications are clinically valid and, if strictly followed, would correspond to good medical practice. For example, this requirement would rule out using 140/90 mm Hg as a complete measure of hypertension control because it ignores the clinical value of lowering blood pressure from 200/140 to 145/95. A measure based on 140/90 also could create an incentive for physicians to overprescribe just to meet the threshold, when settling for, say, 145/95 might be clinically wiser for a particular patient.

*Statistical characteristics.* The second part of the workup is an analysis of a measure's statistical characteristics. This should confirm that the available sample size is sufficient to find clinically important differences in outcomes, that existing information systems can provide the necessary data accurately, and that confounding factors either are not a problem or can be adequately adjusted for. Calculation of the confidence interval around a result should, if possible, include not only the effects of sample size but also the effects of data errors and

confounding. The statistical analysis also should include an assessment of the discriminating power of the measure. Specifically, to what extent do plans vary with respect to the measure? Is this difference a reflection of quality or confounding factors? And to what extent can "low" plans be expected to close the gap?

*Relevance and feasibility.* The analysis of a measure's relevance should describe its clinical importance, economic importance, and comprehensibility. Each of these is made possible by the initial calibration. The analysis of feasibility should consider such things as whether all of the plans have the required sample size, the capabilities of existing information systems, the cost to plans of reporting the measure, and any logistical factors (for example, multiple records, out-of-plan services, and long-term follow-up of patients). Establishing feasibility usually requires field testing.

*Cost-effectiveness.* The formal rationale culminates in an analysis of the measure's cost-effectiveness. This is found by (1) estimating the cost to a plan of collecting data and reporting the measure, (2) estimating the extent to which reporting a measure could possibly cause people to receive better care (for example by causing a "low" plan to raise its performance up to a national benchmark, or by causing people to choose a "higher-performing plan), (3) estimating the health and economic consequences of that degree of improvement in performance, and (4) comparing the cost of measurement to the expected benefits. This information would enable comparisons of the "clinical power" of a measure and would help to ensure that any activity promoted by a measure is in fact worthy of being a strategic goal.

Criteria like these have been described by the NCQA, the Foundation for Accountability (FACCT), and a few other organizations. But two problems remain. First, most organizations that promote measures do not work them up this rigorously. Second, those that try to are severely hampered by the cost. Given the great power of performance measures to influence clinical practice, it is easy to justify that it is worth the cost to determine that they really do what we think they do. Actually coming up with the money will require solving the funding problem.

**Rotation.** The problem of distortion is partially solved by requiring that measures be evidence based, calibrated, and cost-effective. Measures that meet these criteria will represent worthwhile things for plans to do. But the complementary problem—that some plans may deemphasize some things that are not measured—still needs to be solved. One approach is to rotate measures: that is, create and publish a large number of measures (perhaps several hundred) but in any given year use only one or two dozen of them on a rotating basis. Because plans will not know which measures will be chosen, they will need to emphasize all of them. Rotation also will go a long way toward improving the problem of incompleteness. A drawback of rotation is that it complicates serial measurements for monitoring improvements in quality.

**Information Systems.** Ideally, each plan should create an information framework that includes not only an electronic medical record but also the dictionaries, data standards, and linkages required to fully integrate the medical record with other databases. The NCQA has described the specifications for the type of information systems needed to conduct the next generation of performance measures. A temporary alternative to a fully integrated information system is to create computerized registries for specific conditions that are the objects of performance measures.

Information systems also can help with the complexity of health plans. Here the need is not just for an automated medical record but for standardized terminologies and linkages that enable collation of records from all providers. Each plan could require its own automated medical record to be used for its patients but in independent practice association (IPA) models, the practical effect on practitioners, who would have to deal with different systems for different patients in different plans, would be chaos. At the level of individual practitioners, the information systems for each patient should look alike, no matter what plan the patient has. Achieving this will require agreement on standards for such things as a clinical dictionary, data transfer, and user interfaces. Until it occurs, creating a mature performance measurement system for decentralized plans will be either exorbitantly expensive or impossible.

Better information systems also would help to solve the problem of confounding factors. Distinguishing the effects of a plan's actions from the effects of confounding factors requires measuring all of the known factors that can affect an outcome and building a risk-adjustment model that specifies the degree to which the outcome (dependent variable) is affected by each

of the factors (independent variables). Information systems are needed both to build the models and to collect the data needed to apply to the models.

**Case-based measures.** Although each of the previous steps can improve the use of population-based measures, the very nature of these measures limits what they can do. A more complete solution to the problems of bluntness, distortion, incompleteness and cost will require that we eventually supplement population-based measures with what I call case-based measures. Unlike population-based measures, each of which begins with a narrowly defined group of people and asks whether a specific process or outcome occurs, case-based measures begin with a randomly chosen group of people and, through a combination of chart reviews and interviews, asks a variety of questions about their care. Depending on the persons general characteristics (age, sex, risk factors), medical history, and most recent health problems, certain care activities will have been indicated. The actual care delivered can be compared against the indicated activities, as defined by a library of evidence-based guidelines. Because the people interviewed will have a wide variety of medical problems, no particular problem will be represented with sufficient frequency to draw conclusions about specific processes (for example, the rate of eye exams in diabetes). However, the results can be aggregated to provide a far more complete picture of the overall care provided for specific conditions (such as diabetes), populations (such as the elderly), or types of care (such as preventive, chronic, or emergency). The level of detail at which quality can be assessed will depend on the number of cases reviewed and strategies for oversampling. As a rule of thumb, making a statement about the quality of care in a particular area (such as diabetes care) will require about the same number of cases as is required for a single population-based performance measure (such as retinal eye exams).

Case-based measurement complements population-based measurement in several ways. (1) Because it uses charts and patient interviews, it can obtain much more clinical detail than can be obtained for the usual population-based measure. This permits examination of more subtle and clinically relevant aspects of care and goes a long way toward addressing confounding factors. (2) Case-based measures can look at compliance with a very large number of guidelines, without running up huge sample sizes, (3) Through aggregation this approach can make statements about aspects of care that are too small to be seen with population-based measures (for example the management of a collection of rare diseases). (4) Case-based measures can be applied to any type of entity. plans, hospitals, departments. or even individual practitioners. It can work even in complex plans, helping to solve that problem. (5) Because the cases are chosen randomly, a plan must do its best for everyone, and there is no distortion of care. (6) The incentives are correct; to do well, plans should follow evidence-based guidelines. (7) The number of charts needed to assess an HMO is small enough to keep the cost well below the cost of a set of performance measures. (8) The interviews can be used to assess the other dimensions of quality. such as service, choice, and access.

The main drawback of case-based measurement is that it does not measure the use of any particular intervention; it can only make general statements about the quality of care delivered to the population sampled (for example, the members of a plan), or oversampled (for example, diabetics, children). Another important factor is that aggregation requires placing weights on each indicator. A computer-based tool for conducting chart reviews of the type needed for case-based measurement is being developed at RAND.

**Standardized core measurement set.** Eliminating the multiplicity of measurers and measures requires agreement on a single standardized set of measures (a "core measurement set"). Today it would define a set of population-based measures and their specifications, tomorrow it would include the evidence-based guidelines and tools for case-based measures. Having a single core measurement set for the entire country is the only way to identify regional differences, set national benchmarks, compare plans that have national programs or service national corporations. But at a minimum, there must be a single core measurement set for each marketplace. Individual plans can use their own measures for internal quality improvement programs and for monitoring sentinel events, but for comparisons across plans, each plan should only need to report a single core measurement set.

Agreement on a core measurement set would drastically reduce the burden on plans and the confusion of consumers. A single core set also should satisfy the needs of the most purchasers. Purchasers that have additional measurement needs can ask plans to supply more information, but the burden of proof should be on purchasers to conduct the formal work up and document that the measure they want to add is evidence based, calibrated, and cost-effective.

To minimize political influences, the core measurement set ideally should be developed in the private sector. An obvious candidate for a starting point is HEDIS, which was created jointly by plans and purchasers for this purpose and is by far the most widely reported.

**Funding.** The solution to the funding problem is threefold. It is appropriate that plans should bear the cost of collecting data and reporting results, because measuring one's performance is a reasonable part of doing business. In return, plans deserve relief from the multiplicity of measures. Agreement on a core measurement set and requiring a formal rationale for every measure would go a long way toward solving that problem. With regard to the cost of developing, analyzing, and field testing measures, the solution is to acknowledge that performance measurement is a public good and to provide public funding for it. This should be done by the federal government. Even if the cost were \$1 million a measure, it would be trivial compared with the cost of health care in general (about \$1 trillion) and the improvements in quality and efficiency that would follow from better measurement. The last piece of the funding puzzle concerns the cost of administering the core measurement set; selecting and maintaining the measures, publishing the results, and maintaining a database. Based on the NCQA's cost of working up measures and maintaining the Committee on Performance Measurement, this should cost about \$15 million a year. To keep this process as free of political influence as possible, it should be funded by the private sector. The obvious candidates are the purchasers that use the measures. If purchasers cannot agree on a method for covering these costs among themselves, the federal government should do it.

### **Where Do We Go From Here?**

Performance measurement is here to stay. Persistent questions about quality and the tension between quality and cost cannot be resolved without measuring quality. Current efforts to measure performance, as difficult and imperfect as they may be, should be applauded and continued. The problems I have prescribed in this paper are not a reason to stop or slow down; they are a necessary phase in the development of any program to solve a difficult and important social problem. But for performance measurement to move forward, solutions like those just described must be implemented. Here are some steps that might be taken.

Those agencies that are leading the use of measurement sets to accredit plans--such as the NCQA, the joint Commission on Accreditation of Healthcare Organizations (JCAHO), and the American Medical Accreditation Program (AMAP)--should create a core measurement set. The new council just created by those organizations raises hopes that this will occur. The measures for the first version of the core measurement set should be drawn from existing measures for which formal workups have already been done, based on the strengths of the measures as revealed in the workup. Measures that have not yet been formally analyzed should be deferred until their workups can be completed. As soon as the core measurement set has been defined, all plans should report it and all purchasers should accept it. Purchasers that want other measures should do the workups and explain why the additional measures are needed. Plans should be free to ignore requests for additional measures that are not supported by a workup. If necessary, a group such as the new council could serve as a forum for debating the appropriateness of requests for additional measures.

Purchasers and consumer groups should form a foundation to support the council's work. This effort could be initiated by the business groups on health or the Managed Health Care Association. Corporations should contribute according to an appropriate formula. If purchasers fail to provide this support or if "free riders" destroy it, the federal government should provide the funds. Whichever source is used, provision should be made to ensure that the funding is stable and the process is not politicized. The mandate of the council must be to support the most accurate measurement of quality not to promote particular medical activities for particular populations. Nor should it be to protect low-quality plans.

The federal government should provide grants and contracts to support the development, analysis, and field testing of measures. A reasonable amount would be 1 percent of the budget of the National Institutes of Health (NIH)-based on the logic that all the science in the world has no effect until it is implemented properly, and measuring performance is one of the most powerful tools for implementation. The measure development program should be administered by the Agency for Health Care Policy and Research (AHCPR), to ensure its coordination with the agency's centers for evidence-based medicine and other related projects.

Finally, plans and physicians must assume responsibility for improving their information systems so that they can easily get access to the level of clinical detail found in clinical trials. Information systems are extremely expensive, and if they were only useful for performance measurement, they could not be justified. Fortunately, a fully integrated information system is required for other aspects of modern practice of medicine, not just performance measurement.

Performance measurement could revolutionize the practice and cost of medicine. A modest commitment of funds and a reasonable degree of coordination would enable it to reach its full potential.

This paper is based on the Richard and Hinda Rosenthal Lecture, presented by the author at the Institute of Medicine, 30 June 1997. The author has benefited from conversations with many people, especially Beth McGlynn, Margaret O'Kane, Cary Sennett, and members of Kaiser Permanente's Interregional Committee on Performance Measurement.

#### NOTES

1. R.A PauweIs et al., Effect of inhaled Formoterol and Budesonide on Exacerbations of Asthma," New England Journal of Medicine (13 November 1997): 1405.
2. National Committee for Quality Assurance. A Roadmap for Information Systems: Evolving Systems to Support Performance Measurement." HEDIS 3.0/1998. Vol. 4 (Washington: NCQA. 1997)